J Nonverbal Behav (2014) 38:231–245
DOI 10.1007/s10919-014-0173-5

ORIGINAL PAPER

Nonverbal Social Sensing in Action: Unobtrusive Recording and Extracting of Nonverbal Behavior in Social Interactions Illustrated with a Research Example

Denise Frauendorfer · Marianne Schmid Mast · Laurent Nguyen · Daniel Gatica-Perez

Published online: 23 January 2014

Abstract Nonverbal behavior coding is typically conducted by "hand". To remedy this time and resource intensive undertaking, we illustrate how nonverbal social sensing, defined as the automated recording and extracting of nonverbal behavior via ubiquitous social sensing platforms, can be achieved. More precisely, we show how and what kind of nonverbal cues can be extracted and to what extent automated extracted nonverbal cues can be validly obtained with an illustrative research example. In a job interview, the applicant's vocal and visual nonverbal immediacy behavior was automatically sensed and extracted. Results show that the applicant's nonverbal behavior can be validly extracted. Moreover, both visual and vocal applicant nonverbal behavior predict recruiter hiring decision, which is in line with previous findings on manually coded applicant nonverbal behavior. Finally, applicant average turn duration, tempo variation, and gazing best predict recruiter hiring decision. Results and implications of such a nonverbal social sensing for future research are discussed.

Keywords Ubiquitous social sensing platform · Automated extraction · Applicant nonverbal behavior · Hiring decision · Job interview

Introduction

Observing and coding of nonverbal behavior in social interactions is time and resource intensive. Although technical advances in recording devices make it relatively easy to record nonverbal behavior during social interactions with regular video cameras and

D. Frauendorfer (✉) · M. Schmid Mast
Department of Work and Organizational Psychology, University of Neuchatel, Emile-Argand 11, 2000 Neuchatel, Switzerland
e-mail: denise.frauendorfer@unine.ch

L. Nguyen · D. Gatica-Perez
Idiap Research Institute, Ecole Polytechnique Federal de Lausanne (EPFL), Lausanne, Switzerland

microphones, the extraction of nonverbal cues from the recordings still typically needs to be done by humans who manually code each and every behavior of the interaction partners (Reis and Charles 2000).

One way to circumvent the tedious coding is to use wearable sensing devices that measure the nonverbal behavior as it occurs from the point of view of the social interaction partner. An example of this would be eye tracking devices mounted on the head of the study participant or body-worn sensors to detect gestures (Junker et al. 2008). These wearable sensing devices have the disadvantage that they interfere—at least to some extent—with the social interaction per se and sometimes with the very nonverbal behavior that is measured (e.g., a worn eye tracking device can affect a person's gazing or how an interaction partner gazes at the person). If social interaction behavior could be observed without having to put on cumbersome sensing devices on participants and, if researchers (and practitioners) had an account of the social interaction partners' nonverbal behavior available right at the end of a social interaction, without having human coders invest many hours to code these behaviors, this could immensely propel the nonverbal behavior field.

In the present article, we will show how far along we have come in this endeavor through collaboration between psychologists and computer scientists with what we call *nonverbal social sensing* (described in more detail below). Nonverbal social sensing can be used to unobtrusively record and extract nonverbal cues in social interactions. We will present a nonverbal social sensing platform, a smart room that we created for the study of nonverbal (vocal and visual) conversation behavior in dyadic interactions.

The goal of the present paper is to show how nonverbal social sensing can be used for research in psychology by providing a concrete research example from the field of personnel psychology (i.e., the link between applicant nonverbal behavior in the job interview and the recruiter's hiring decision). We set out to test whether visual and vocal nonverbal cues obtained via nonverbal social sensing can be used to replicate findings in the psychological literature.

Nonverbal Social Sensing

Nonverbal social sensing means the sensing and recording of nonverbal cues from people engaged in social interactions via ubiquitous computing platforms and computational models and algorithms for extracting those cues. Ubiquitous computing means that the user does not need to enter the computer environment but that the computer fits the human environment. Social sensing and recording devices are incorporated in the everyday environment of a person; the environment becomes "smart". There are stationary sensing and recording platforms (e.g., "smart rooms") and wearable sensing and recording platforms (e.g., smartphones). In this paper, we focus on stationary social sensing platforms, which are smart rooms equipped with sensors (e.g., cameras, microphones, Kinect motion sensor) that record the nonverbal behavior of the social interaction partners in the room without them wearing any special gear or sensors. We focus on the sensing and the extraction of nonverbal cues and include visual nonverbal cues (e.g., smiling, gazing, and nodding) as well as vocal nonverbal cues (e.g., audio back-channeling, average turn duration, and tempo variation).

Nonverbal social sensing is based on two steps. First, the social interaction behavior is sensed and recorded in an unobtrusive way, via the ubiquitous sensing platform. Second, the nonverbal behavior is extracted based on the recordings with computational models and algorithms (Gatica-Perez et al. 2007; Vinciarelli et al. 2008). Once those models and algorithms are stable and accurate, the nonverbal cue extraction becomes automated.

Unobtrusive Sensing

Due to the non-invasive nature of ubiquitous computing, there is no interference of the recording with the social interaction and the sensing devices do not restrict the social interaction partners in any way during the conversation. Especially when studying nonverbal behavior, the unobtrusiveness of the sensing and recording device is important. As an example, if a researcher is interested in measuring gesturing and the social interaction partners are equipped with body-worn sensors, this can impede spontaneous gesturing because the placement of the sensors might make certain movements harder to do, or because the sensors feel heavy, or because the person wearing the sensors is conscious about the fact that his/her gesturing behavior is surveyed.

An example of a stationary social sensing platform, a smart room, can be seen in Fig. 1. This is the nonverbal social sensing platform with which we have recorded the data presented in this article. The platform is set up to record two interaction partners (in our case, the job applicant and the recruiter) with two HD cameras and a microphone array (i.e., "microcone") recording the interaction partners' voices separately (i.e., direct speaker segmentation). The "microcone" (http://www.dev-audio.com/products/microcone/) consists of a microphone array able to register up to six people at the same time and each person is recognized individually and his/her speech registered separately from the others. This so-called direct speaker segmentation is a key advantage for the automated extraction of the vocal nonverbal cues. The cues are directly available for each interaction partner separately while at the same time they remain synchronized among the different interaction partners to, for instance, easily extract interruptions or other turn taking behavior. When putting one microphone in the room to register different interaction partners, the speaker segmentation has to be conducted as an additional step (Lathoud and McCowan 2003). When clipping an individual microphone on each interaction partner, the synchronization of the different recordings needs an additional step in order to be able to extract turn taking behavior. Moreover, individual microphones clipped to the social interaction partners make the behavioral observation salient and less unobtrusive. Using ubiquitous computing like in the form of the "microcone" makes it possible to take the sensors, the microphones, off the individual interaction partners and put them in the environment to make the room smart.

Our nonverbal social sensing platform has been created to sense and register nonverbal communication behavior in dyadic interactions between two social interaction partners that sit at a table facing each other. This is a common setting in the real world (e.g., job interview, physician–patient communication). It goes without saying that there are challenges inherent in using a smart room like ours for nonverbal social sensing. The smart-room environment needs to fulfill certain criteria in terms of camera resolution, lighting conditions, and camera angles for the accurate automated extraction of nonverbal cues. For camera resolution we used 1280 × 960 cameras, which allowed the recording of detailed images; however, lower resolution (e.g., 640 × 480) video cameras can also be used as some automated visual nonverbal cue extraction method do not require very high video resolution. The lighting conditions of the room are critical for the proper recording of interactions and can be overcome by using blinds or curtains in front of the windows (to prevent reflections and ensure constant lighting conditions) and using artificial lighting. In terms of camera angles, the devices should be placed in a manner such that the views are quasi-frontal, while avoiding attracting the participants' attention; pilot-studies are necessary to fine-tune the sensor setting. The sensing platform also has requirement in terms of storage. Typically, a 10-min interaction requires 2 GB of storage while recorded at full

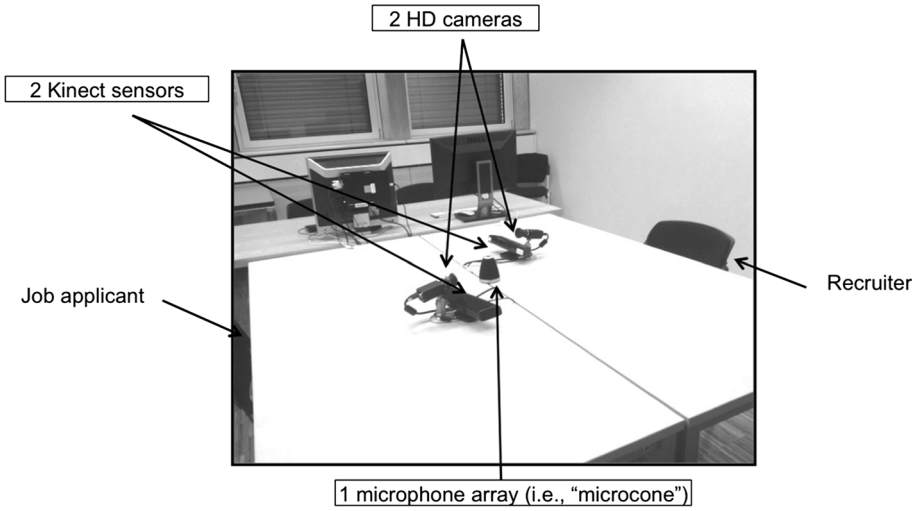

Fig. 1 The stationary indoor sensing platform with two high definition cameras, two Kinect sensors, and the microphone array ("microcone")

resolution. This relatively large data size can be reduced by adjusting the resolution of the video cameras. Furthermore, with the dramatic increase of storage capabilities, the dataset storage size becomes less and less a problem.

Automated Extraction

The automated extraction of the sensed and recorded interaction behavior requires the development of algorithms and machine learning technologies by computer scientists. They train computers to extract the nonverbal cues from the sensed and recorded data. The automated extraction of nonverbal cues circumvents the hand coding by humans and offers therefore a tremendous advantage for nonverbal behavior researchers. Once the algorithms are in place and validated, the coding can be done in an automated way by computers. This means that the researcher can investigate nonverbal behavior in social interactions with many social interaction partners present at the same time and over long recording periods. Because the coding is automated, the amount of data processed does not make a difference any more.

In the domain of nonverbal cue extraction, some of these automated methods have been demonstrated to be accurate and have the advantage of being publicly available, such as voice energy, pitch, and speaking rate (Basu 2002), visual motion (Biel et al. 2011), or head pose extraction (Ba and Odobez 2011). Other cues can be directly extracted from the recorded data of specific ubiquitous sensing devices such as, for instance the "microcone", for which pauses, total speaking time, or interruptions can easily be extracted. Some nonverbal cues are more challenging to extract from videotaped material. Cues like arm (Marcos-Ramiro et al. 2013) and head gestures (Nguyen et al. 2012), or gaze direction (Funes and Odobez 2012) are still subject of ongoing perceptual computing research with promising preliminary results.

It has to be noted that the ubiquitous sensing and the automated extraction processes go hand in hand. This means that the algorithms developed based on a specific social sensing

platform work best for data that has been collected in this environment. As an example, if the recording angle changes, the algorithm that has learned to detect gazing from videos showing the social interaction partners at a specific angle, will not work as well or not at all. If the algorithm has learned to accurately detect gazing when social interaction partners have been filmed front view, the same algorithm will not be able to detect gaze if the social interaction partner is filmed profile view.

At least for the extraction of visual cues such as nodding, stationary social sensing platforms are an advantage because the sensing devices remain at the same place. It becomes, however, also clear that there are some constraints on the social interaction partners necessary to optimize automated nonverbal cue extraction. In the case of our smart room, social interaction partners need to sit at a table facing each other. Given that such a scenario is very common in our everyday lives (e.g., job interview), the constraints become minor.

We have argued that using nonverbal social sensing brings about two distinct advantages for nonverbal behavior researchers. First, the unobtrusiveness of the recording of the social interaction makes it possible to obtain nonverbal behavior information as it occurs in social interactions without in any way interfering or disrupting the interaction. Second, the automated extraction of the nonverbal cues makes it possible to study the nonverbal behavior of many social interaction partners over long periods of time without this necessitating costly and time consuming human coders. In principle, once the nonverbal social sensing is in place, the researcher has the nonverbal behavior data of each of the social interaction partners at hand right away.

Nonverbal Social Sensing in Action

We used nonverbal social sensing to investigate how nonverbal job applicant behavior is related to the recruiter's hiring decision. Research shows that nonverbal behavior of an applicant during the job interview indeed influences the hiring decision of the recruiter. For instance, applicants who use a high amount of eye contact, smiling, body orientation toward the interviewer, and less interpersonal distance are perceived as being more hirable, more competent, more motivated, and more successful than applicants who do not express such immediacy behavior (Imada and Hakel 1977). Applicants with extended eye contact, a loud and modulated voice, fluent speech, and an expressive face (expression of affects) during the job interview were more likely invited for a second job interview than applicants who did not show these behaviors (McGovern and Tinsley 1978). Forbes and Jackson (1980) showed that employed applicants looked more at the recruiter, smiled more, and nodded more during the job interview than applicants who were not employed after the job interview. Also, Anderson and Shackleton (1990) report that selected applicants showed more gazing and more facial expressiveness than rejected applicants. In sum, research shows that the more an applicant shows immediacy behavior the better his or her chances of being hired. The immediacy behavior hypothesis claims that immediacy nonverbal behavior of the target (e.g., applicant) is perceived as closeness and involvement in the interaction (Mehrabian 1972), which in turn is perceived by the observer (e.g., recruiter) in a positive way affecting his or her evaluation of the target. We investigate whether similar results can be achieved with unobtrusive sensing and automated extraction of the applicant's nonverbal behavior during a job interview.

Some of the research on job applicant nonverbal behavior and hiring decision relies on the investigation of single cues (does more smiling increase the likelihood of being hired?)

(Anderson and Shackleton 1990; Forbes and Jackson 1980) and other research has looked at a behavioral composite, typically lumping together many different cues and investigating their joint effect on the hiring decision (does more smiling, more gazing, and more nodding increase the chances of being hired?) (Imada and Hakel 1977; McGovern and Tinsley 1978). What is not clear is whether different cues together, how many of them, and which combination of those cues best predict hiring success. This is the reason why we investigated which *combination* of applicant nonverbal cues predicts the hiring decision best. For instance, is it more the applicant's gazing in combination with other visual nonverbal cues such as smiling and nodding, or is it the applicant's gazing in combination with vocal nonverbal cues, such as audio back-channeling (e.g., "mmhh", "yes"), average turn duration (i.e., average length of turns), and tempo variation (i.e., the extent to which the applicant varies in his or her speech tempo) that predict hiring decision?

To decide which nonverbal behaviors to extract and focus on, we selected applicant nonverbal immediacy cues shown in previous research as being related to the hiring decision (e.g., Anderson and Shackleton 1990; Forbes and Jackson 1980; Hollandsworth et al. 1979; Imada and Hakel 1977; McGovern and Tinsley 1978). For the visual nonverbal behavior we selected three cues being indicative of immediacy behavior: smiling, gazing, and visual back-channeling (i.e., nodding while listening to the interaction partner). For applicant vocal nonverbal behavior, we selected the following three immediacy cues: audio back-channeling (i.e., short utterances like "mmhh", "ah", or "yes" while listening to the recruiter), average turn duration (as an indicator of speech fluency, Fillmore 1979), and tempo variation. So far, tempo variation has been neglected in the nonverbal immediacy behavior research. However, it has been clearly defined by Guerrero (2005) as being an important indicator of immediacy, as have been the other above mentioned visual and vocal nonverbal cues. Thus, we expect applicant tempo variation to be positively related to hiring decision.

In sum, the research questions we pursue in the present work are: (a) Can applicant visual and vocal nonverbal immediacy behaviors validly be obtained via nonverbal social sensing? (b) Are the so obtained applicant nonverbal behaviors linked to better hireability judgments of recruiters? (c) Which combination of the applicant's nonverbal behaviors best predicts hiring? By answering these research questions we can demonstrate the validity of the nonverbal social sensing approach in two ways: First, by showing the validity of the automated extraction process when compared to hand coding by humans, and second, by replicating findings established in the literature when using the cues that have been extracted by nonverbal social sensing. We hope to provide evidence showing that nonverbal social sensing is a powerful tool that can be used by nonverbal behavior researchers.

Method

Participants

Sixty-two participants (45 female and 17 male) with a mean age of 24 years (one missing; range from 19 to 40 years) applied for a sales job. Participants were Bachelor and Master students majoring in different domains (90 %), Ph.D. students (4.8 %), and employed (3.2 %; one missing). We posted a job advertisement at different places, such as online forums and information boards at different universities.

Procedure

The job advertisement said that we are looking for research assistants to help us recruit community people on the street as participants for studies conducted at the local university. The advertisement specified the competences required for the job (communication and persuasion skills, extraversion, and agreeableness). We also included the exact application procedure and the e-mail address of the first author, so that participants could get in contact. People interested in the job were invited for an assessment including a structured job interview.

Participants coming to our lab were welcomed by the recruiter (first author) and given an informed consent form to sign. Participants gave their written consent to being video and audio taped during the job interview in order to use this data for scientific purposes. They underwent a structured job interview (average interview duration: 11 min). The job interview contained introductory questions (e.g., "could you say a few words about yourself?", "what is your motivation to apply for this job?") and behaviorally based questions (e.g., Campion et al. 1994; Janz 1982; Motowidlo et al. 1992) concerning the competences required (e.g., for communication skills: "Could you tell me about a situation in which you succeeded in communicating in a clear and well-structured way?" and for persuasion: "Could you tell me about a situation in which you succeeded in persuading another person of something?"). As a last question of the interview, applicants were asked to tell the recruiter about their strengths and weaknesses. At the end of the interview, participants were informed that they would be hired for the job. All applicants received the same information without them knowing that we told everybody that he/she was hired. Based on the video and audio recordings, we extracted and coded the applicants' nonverbal behavior and we had professional recruiters watching the videotapes (with sound) and provide us with a hiring decision for each applicant.

Material and Measures

Hiring Decision

We contacted professional recruiters via e-mail asking them to evaluate the videotapes showing applicants during the job interview. Recruiters were offered a salary of an equivalent of \$100/h. Five recruiters (2 male, 3 female, mean age: 26 years, $SD = 7.67$, mean experience: 5.7 years) agreed to do the evaluation. All applicants were evaluated by three recruiters (see inter-rater correlation below). That is, one recruiter evaluated all applicants and four recruiters each viewed half of the job interviews. We provided recruiters with a detailed job description. Recruiters evaluated applicants with respect to hireability on a scale from 0 to 100 % ("to what extent would you think that this applicant should be hired for the job"), $M = 59.73$, $SD = 19.15$, $ICC[1] = .50$, $ICC[2] = .75$, $F = 4.45$, $p < .05$.

Applicant Nonverbal Behavior

Our smart room (Fig. 1) is equipped with two high definition cameras recording video at 26.6 frames per second and a commercial microphone array device ("microcone"). The "microcone" (http://www.dev-audio.com/products/microcone/) is a microphone array that registers up to six people at the same time. The software coming with this device provides automatic speaker segmentation based on spatial discrimination. The resulting

segmentations are stored in a file containing the identifier (recruiter or job applicant) and the relative time in seconds (starting and ending time of the speaking segment). We used the segmentation files to extract on the one hand the vocal audio signals for each interaction partner (i.e., the applicant's tempo variation) and on the other hand the speaking turn-based nonverbal cues: the average time of applicant's turns and the number of short utterances like "mmhh", "ah", or "yes" while listening to the recruiter as (audio back-channeling). The three vocal nonverbal cues were thus obtained by automated extraction. For the visual nonverbal cues, we gained applicant visual back-channeling via automated extraction (described in more detail below) while we hand coded gaze and smiling. The work on automated extraction of gaze and smiling is still in progress which is why the applicant's smiling and gazing behavior were coded manually.

Applicant Audio Back-Channeling Applicant audio back-channeling was defined as events when the applicant produces a short utterance (such as "mmhh" or "yes") while the recruiter is speaking. Concretely, short utterances were defined as speaking segments of duration shorter than 1 s, and we recorded the number of such events. This number was then normalized by the job interview duration, $M = 6.95$, $SD = 5.86$.

We had two human raters do hand coding of the applicant's audio back-channeling ("count the number of times an applicant says "yes" or "mmhh" or another short—shorter than 1 s—vocalization of approval or agreement while the applicant is listening to the recruiter") based on the videotapes for 20 job applicants. Inter-rater reliability of the two coders was $r = .94$. We averaged the raters' codings.

Applicant Average Turn Duration The average time of speaking turns of the applicant was assessed using the speaking segmentations. We aggregated the total speaking time, and divided this number by the total number of speaking turns (without taking into account the utterances shorter than 1 s), $M = 30.77$, $SD = 13.94$.

The computation of the average turn duration relies on the speaker segmentations (when is the applicant speaking). Speaking turns are characterized by their starting and ending time and are assessed based on two pre-defined conditions. First, short utterances (speaking turns shorter than 1 s) are discarded. Second, if the non-speaking segment between two turns is shorter than 1 s, the two turns are merged, being considered as one turn. In contrast, the turns are considered as two separate turns, when the non-speaking segment between the turns is longer than 1 s. Given this highly technical assessment, we assume the average turn duration to be a technical rather than a psychological measure. The assessment of the average turn duration is based therefore solely on the automated extraction. Notwithstanding that average turn duration can affect observer judgments without observers being able to explicitly extract such a "technical" cue. It can affect judgment on an implicit level, maybe in combination with other cues.

The average time of speaking is defined as being part of speech fluency. A person who talks fluently succeeds in talking at length (Fillmore 1979). In other words, the longer the average time of the speaker's turn is, the more he or she talks fluently.

Applicant Tempo Variation To assess the applicant's tempo variation, we used a hidden Markov model (HMM) and Mel-frequency cepstrum coefficients (MFCC), using the method presented in Basu (2002). To detect the voice segments we used open-source matlab code provided by the MIT Media Lab (http://groupmedia.media.mit.edu/data.php). The tempo (or speaking rate) was defined as the number of voiced segments (i.e., vowels)

per second. We computed the standard deviation of the number of vowels per second across the speaking time, $M = 1.25$, $SD = .09$.

Research shows that the just noticeable difference in speaking rate is 5 % deviated from the baseline (Quené 2007). Given this, we assume this nonverbal cue to be a technical rather than a psychological measure. This is why the assessment of this nonverbal cue was based solely on the automated extraction.

Applicant Visual Back-Channeling Applicant visual back-channeling was defined as an applicant's head nod while the recruiter was speaking. We used the method presented in Nguyen et al. (2012) to automatically detect head nods from the videos, providing the relative timings of nodding events (starting and ending time). This method starts by tracking the location the applicant's face and estimates the visual motion in the face region. The horizontal and vertical visual motions are then analyzed in terms of their frequency, using a windowed Fourier transform, and the binary decision between a nod and a non-nod is performed using a standard machine learning model. As head gesture dynamics are conditioned by the speaking status of the person under analysis, we used two support vector machine classifiers (one for speaking and one for silent, provided by the "micro-cone" segmentations) to improve the accuracy of the detection. The objective evaluation of the nodding extraction method was presented in Nguyen et al. (2012) where the F1-score (harmonic mean between precision and recall) was 63 % at the frame level. Even if the method is not 100 % accurate, the detection is satisfactory for detecting most head nods (even subtle ones), while keeping the number of false detections low. Moreover, as the detected nods are aggregated over the whole job interview, the errors tend to cancel out. The nodding extraction method is completely repeatable, meaning that if a same video is processed several times, the detected nods will always be the same at each run. The head nod detection outputs the number of nodding events. The number of head nods was normalized by the job interview duration, $M = 21.37$, $SD = 14.15$.

We had two human raters do hand coding of the applicant's nodding while listening (visual back-channeling: count the number of times an applicant nods while the applicant is listening to the recruiter) based on the videotapes for 20 job applicants. Inter-rater reliability of the two coders was $r = .90$. We averaged the raters' codes.

Applicant Smiling A trained rater coded the degree to which an applicant smiled during the job interview as a general impression every minute of the job interview on a scale from 1 (*not at all*) to 5 (*very much*). Applicant's smiling was defined as the edges of the lips moving upwards compared to the person's baseline lip expression. Raters viewed a couple of min of the applicant in the job interview to get an idea about that particular applicant's baseline lip expression. The general impression score was computed based on the average across all 1-min thin slices, $M = 2.51$, $SD = .90$. A second rater coded five applicants to obtain an inter-rater reliability measure ($r = .95$).

Applicant Gazing Gazing was defined and coded as the applicant looking at the recruiter. The duration of applicant gazing was divided by the job interview duration to obtain the applicant gazing measure, $M = .51$, $SD = .14$. A second rater coded five applicants to obtain an inter-rater reliability measure ($r = .95$).

Results

To know whether the applicant's visual and audio back-channeling can be obtained validly via nonverbal social sensing, we compared this nonverbal behavioral data retrieved from the automated extraction with the manual coding conducted on a subsample of 20 job interviews. Automatically extracted visual back-channeling and audio back-channeling correlate significantly with the manual coding of these nonverbal cues, $r = .88$ for visual back-channeling, $r = .78$ for audio back-channeling, respectively.

To test whether the job applicant's nonverbal behavior during the job interview is linked to better hireability judgments of recruiters, we conducted a multiple regression analysis in which we regressed hiring decision on the applicant's nonverbal behavior (i.e., visual nonverbal behavior: smiling, gazing, visual back-channeling and vocal nonverbal behavior: audio back-channeling, average turn duration, and tempo variation), controlled for applicant gender and age.

The regression model was significant, showing that applicant nonverbal behavior significantly predicts hiring decision (Table 1). Job applicants were more likely to be hired when *gazing more* at the recruiter, having *longer average speaking time turns,* and when *varying their tempo more* during the job interview. When calculating separate regression models for vocal and visual nonverbal cues, results show that both regression models are significant meaning that visual cues alone (Table 2) and vocal cues alone (Table 3) both predict hiring decision significantly.

To investigate which combination of the applicant nonverbal behavior cues best predicts hiring, we performed a sequential forward search (SFS) (Pudil et al. 1994) on the six selected cues, which is a standard feature selection technique in machine learning to find the best combination of cues with respect to a merit function (i.e., the function we want to maximize). To this end, we used the *R*-square value as the merit function in a multiple linear regression framework. SFS searches the feature space (in this case, we have six nonverbal cues, therefore it is a six-dimensional space) for the best combination of features using a greedy approach which (a) starts by selecting the best (with respect to the merit function) feature alone (b) then tries all combinations between the selected feature and the remaining ones and adds the best one to the selected set. The process is repeated until the maximum number of features (a parameter set by the user) is reached or until the merit function decreases. In the present study, the maximum number of features allowed was set to 6, meaning that if necessary, all features could be used to predict the hiring decision, if that would increase the merit function.

With this method, the best combination of cues is the *average turn duration* combined with the *tempo variation*, and *gazing,* yielding a R^2 of .38. Taken individually, the *average turn duration* is the most predictive nonverbal cue ($R^2 = .16$), followed by *gazing* ($R^2 = .11$) and *tempo variation* ($R^2 = .02$). Adding more nonverbal cues decreases the prediction performance.

Discussion

The aim of the current article was to introduce nonverbal social sensing as a powerful research tool for nonverbal behavior researchers. We tested whether applicant visual and vocal nonverbal immediacy cues obtained via nonverbal social sensing are valid predictors of being hired. We also asked which combination of the applicant nonverbal behaviors best predicts hiring.

Table 1 Multiple regression analyses with applicant nonverbal behavior predicting hiring decision (controlled for applicant gender and age)

Applicant variable	B (*SE* B)	β
Gender	−3.37 (5.41)	−.08
Age	−.78 (.56)	−.15
Smiling	−1.93 (2.51)	−.09
Gazing	56.08 (16.00)	.40**
Visual back-channeling	.05 (.16)	.03
Audio back-channeling	−.09 (.39)	−.03
Average turn duration	.66 (.14)	.48***
Tempo variation	61.79 (23.67)	.28*

$R^2 = .53$ ($p < .0001$); * $p < .05$; ** $p \leq .001$; *** $p < .0001$. $F = 7.17$, $df = 60$

Table 2 Multiple regression analyses with applicant visual nonverbal behavior predicting hiring decision (controlled for applicant gender and age)

Applicant variable	B (*SE* B)	β
Gender	−10.08 (5.95)	$-.24^{+}$
Age	−.60 (.62)	−.12
Smiling	−4.32 (3.00)	−.20
Gazing	65.05 (18.01)	.46*
Visual back-channeling	.08 (.18)	.06

$R^2 = .27$ ($p = .003$); $^{+}$ $p < .10$; * $p \leq .001$. $F = 4.11$, $df = 60$

Table 3 Multiple regression analyses with applicant vocal nonverbal behavior predicting hiring decision (controlled for applicant gender and age)

Applicant variable	B (*SE* B)	β
Gender	−5.16 (4.72)	−.12
Age	−.92 (.58)	−.18
Audio back-channeling	.46 (.36)	.14
Average turn duration	.74 (.15)	.54**
Tempo variation	57.82 (25.45)	.26*

$R^2 = .41$ ($p < .0001$); * $p < .05$; ** $p < .0001$. $F = 7.59$, $df = 60$

We validated the nonverbal social sensing method by correlating some of the automatically extracted nonverbal behavior data with nonverbal behavior data coming from human coders. Results indicate that social sensing for the applicant's visual and audio back-channeling is highly valid, which is promising news. In other words, nonverbal social sensing extracting the applicant's visual and audio back-channeling can therefore be used to replace human coders in the future, which in turn lowers costs and time investment. Moreover, the present study shows that the applicant's average turn duration and tempo variation predicts recruiter hiring decision. In other words, the applicant's average turn duration, an indicator of speech fluency (Fillmore 1979), can be seen as being validly extracted to the extent that it replicates previous research, which coded speech fluency manually (Hollandsworth et al. 1979). In terms of the applicant's tempo variation, research has not yet investigated how it is related to hiring decision. However, because Guerrero (2005) claims that tempo variation is an indicator of immediacy nonverbal behavior, which in turn is known to predict hiring decisions, our results show first evidence confirming Guerrero's argument and demonstrating that tempo variation can be validly extracted.

With a study in the field of work and organizational psychology we showed that automatically extracted applicant nonverbal immediacy behavior predicts outcomes such as hiring decision. Results indicate that applicant nonverbal behavior during the job

interview significantly affects the hiring decision such that applicants who had longer average turn duration, varied their speech tempo more during the job interview, and gazed longer at the recruiter, were more likely to be hired than applicants using those nonverbal cues to a lesser extent. These findings confirm previous research and with that the immediacy hypothesis (Mehrabian 1972), stating that the more the applicant shows immediacy nonverbal behavior the more likely the applicant is hired. The novelty of these results lies in the fact that the computational nonverbal behavior analysis leads to similar results compared to previous research based on manually coded nonverbal behavior (Anderson and Shackleton 1990; Forbes and Jackson 1980; Imada and Hakel 1977; McGovern and Tinsley 1978). This adds evidence to the validity of the nonverbal social sensing method.

Moreover, we were interested in a rather novel research question in the field of nonverbal communication: Which *combination* of the applicant's nonverbal behavior cues *best* predicts the hiring decision. In other words, how many and which applicant nonverbal behaviors best predict the recruiter's hiring decision? Results reveal that the combination of the applicant's average turn duration, his or her increased tempo variation, and maintenance of eye contact best predict hiring decision. Among the three cues, applicant average turn duration is the more predictive one.

Fillmore (1979) defined speech fluency among other things as talking at length with making few pauses. Research shows that speech fluency has a remarkable impact on persuasion and related constructs in sales. For instance, Burgoon et al. (1990) found that speech fluency is the strongest predictor for perceived competence and credibility and it is one of the strongest predictors for persuasiveness. Moreover, speech fluency is associated with trustworthiness, dynamism, and attitude change in the context of sales (Leathers 1992; Leigh and Summers 2002). Leigh and Summers (2002) argue that through speech fluency, assertiveness, and control is communicated (Burgoon 1994; Wardhaugh 1985) and thus, a higher likelihood of persuading the client is achieved.

In terms of tempo variation, Woolbert (1920) showed that the higher the tempo variation, the higher the audience retention is. Moreover, Knapp and Hall (2010) argue that the higher the variability in speech tempo, the better the message delivery, compared to speakers holding the speech tempo constant. Maintaining eye contact with the social interaction partner results in perceiving the communicator as more sincere, honest, competent, credible, self-confident, dominant, and likeable (Kleinke 1986). Moreover, sales people's gazing at the client has shown to predict how they are perceived in terms of warmth and trustworthiness, which are two key aspects in sales (Leigh and Summers 2002; Wood 2006). Thus, in the present study, we suspect that the recruiters inferred high persuasiveness, credibility, and a high ability to deliver information efficiently and recognized these as important aspects for the job which is why they recommended hiring those applicants. The job for which the applicants were recruited was a job necessitating those competencies (convince people on the street to participate in studies at the local university). Alternatively, the recruiters might themselves have been persuaded to hire the applicants by the applicant's immediacy behavior without necessarily thinking about the job competencies.

Although the literature shows that immediacy behavior of the applicant is a plus for applicant evaluation for many different jobs, the here reported results might differ for different types of jobs. For jobs requiring less social interaction, the nonverbal behavior of the job applicant might influence the hiring decision less because research shows that nonverbal behavior of the job holder is important, especially for jobs with extended client contact (Leigh and Summers 2002; Peterson 2005; Taute et al. 2011; Wood 2006). In the

present study, applicants applied for a job similar to sales. Thus, for evaluators, the ability to speak fluently and to vary in tempo variation to keep the attention of potential clients while talking to them seems an important factor as detailed above. However, as job profiles differ across domains, the applicant's nonverbal behavior profile related to better chances of being hired might also differ. Depending on the type of job for which applicants are applying, different nonverbal cues and different combinations of nonverbal behaviors might predict hiring decision.

We always used the same trained recruiter who interviewed all applicants and behaved in a standardized way. It is therefore unlikely that the applicant's behavior was just an expression of the recruiter's behavior (i.e., mimicking the recruiter's immediacy behavior; Tiedens and Fragale 2003). In order to rule out this possibility, we also coded some of the same vocal and visual behaviors of the recruiter and for each of the behaviors we correlated the recruiter's with the applicant's behavior. Correlations were small and non-significant (ranging from $r = -.14$ to $r = .16$, all p's $> .20$). Although we cannot completely rule out that some of the applicant's behavior was a reaction to the recruiter's behavior, the non-existence of mimicry with respect to the behaviors we tested, provides some support for the claim that the behavior we measured in the applicant originated in the applicant.

We showed that nonverbal social sensing worked: we validated some of the automated cue extraction with hand coding and we validated the entire method with replicating research findings in the literature. Moreover, we also foreshadow that the collaboration with computer scientists can open up new and promising future avenues of research when testing which cue combinations best predict hiring. Nonverbal social sensing can be the basis of further interesting analyses: the extraction of nonverbal composites. Much research on nonverbal behavior relies on single nonverbal cues. However, nonverbal composites—a specific combination of cues—might be a more promising indicator of interaction outcomes. As an example, one of the strongest and most stable predictors of dominance and status in a dyadic interaction is visual dominance. The visual dominance ratio (VDR) (Dovidio et al. 1988) is defined as the percentage of gazing at an interaction partner while speaking divided by the percentage of gazing while listening. A human coder cannot code this indicator in one coding because there are too many cues to pay attention to. So several rounds of coding are necessary. With nonverbal social sensing, the coding of behavioral composites becomes easy. The same is true for the question that we pursued in the current paper, namely which combination of nonverbal cues best predicts an outcome.

We have focused on a stationary ubiquitous computing platform in the current paper. Ubiquitous mobile social sensing (with smartphones) is another interesting research avenue for nonverbal social sensing. While the extraction of visual cues in ubiquitous mobile sensing is difficult, the vocal nonverbal cues can be sensed and extracted with relative ease. As an example, with other collaborators we investigated a method that infers felt stress by the smartphone user based on nonverbal social sensing of the user (Lu et al. 2012).

The setup of the stationary nonverbal social sensing platform we used requires costs, time, and knowledge to deal with different factors, such as light conditions (i.e., artificial light vs. day light), possible reflections coming from the window, resolution of the cameras, and high capacity of the computer so that registering the video while recording is enabled. Moreover, the distance of the cameras to the person being registered and the camera angles are important because the registered person has to appear in the picture frame all the time with no body parts cut-off when moving too extensively. Once the sensing platform is in place, it can be used repeatedly, of course. Note that algorithms are mostly calibrated to the sensing devices. Only few algorithms are commercially available,

such as the "microcone" which is a more consumer-oriented device in our setting. This device has a standard way of use and can therefore be easily installed and used.

Nonverbal social sensing is not only an interesting research tool. It can also be interesting for practitioners, for instance, in the realm of personnel selection as we have shown with our data. If the computer can approximate the hiring decision of the recruiter, maybe the recruiter can be assisted in his or her decision by nonverbal social sensing. Given that research shows that recruiters often are biased in their decision making by applicant characteristics that are not central to the job at hand (e.g., gender, appearance, stereotypes) (e.g., Anderson 1992; Barrick et al. 2009), automated nonverbal sensing has the potential to help overcome such biases.

Acknowledgments We thank Dr. Florent Monay (Idiap) for the design and implementation of the sensing platform; Dr. Jean-Marc Odobez (Idiap) for his contribution to the sensing platform and the nodding recognition method; and Prof. Tanzeem Choudhury (Cornell University) for her contribution to the design of the job performance part of the study. This research was funded by the Swiss National Science Foundation through the Sinergia SONVB (Sensing and Analyzing Nonverbal Organizational Behavior) project.

Conflict of interest The authors declare that they have no conflict of interest.

References

Anderson, N. R. (1992). Eight decades of employment interview research: A retrospective meta-review and prospective commentary. *European Work and Organizational Psychologist, 2*, 1–32. doi:10.1080/09602009208408532.

Anderson, N. R., & Shackleton, V. J. (1990). Decision making in the graduate selection interview: A field study. *Journal of Occupational Psychology, 63*, 63–76. doi:10.1111/j.2044-8325.1990.tb00510.x.

Ba, S., & Odobez, J.-M. (2011). Multiperson visual focus of attention from head pose and meeting contextual cues. *IEEE Transaction on Pattern Analysis and Machine Intelligence, 3*, 101–116.

Barrick, M. R., Shaffer, J. A., & DeGrassi, S. W. (2009). What you see may not be what you get: Relationships among self-presentations tactics and ratings of interview and job performance. *Journal of Applied Psychology, 94*, 1394–1411.

Basu, S. (2002). *Conversational scene analysis.* MIT Department of EECS, Cambridge, MA. http://alumni.media.mit.edu/~sbasu/papers.html.

Biel, J.-I., Aran, O., & Gatica-Perez, D. (2011). *You are known by how you vlog: Personality impressions and nonverbal behavior in YouTube.* Paper presented at the proceedings of international AAAI conference on weblogs and social media, Barcelona, Spain.

Burgoon, J. K. (1994). Nonverbal signals. In M. L. Knapp & G. R. Miller (Eds.), *Handbook of interpersonal communication* (pp. 229–285). Thousand Oaks, CA: Sage.

Burgoon, J. K., Birk, T., & Pfau, M. (1990). Nonverbal behaviors, persuasion, and credibility. *Human Communication Research, 17*, 140–169.

Campion, M. A., Cheraskin, L., & Stevens, M. J. (1994). Career-related antecedents and outcomes of job rotation. *Academy of Management Journal, 37*, 1518–1542.

Dovidio, J. F., Brown, C. E., Heltman, K., Ellyson, S. L., & Keating, C. F. (1988). Power displays between women and men in discussions of gender-linked tasks: A multichannel study. *Journal of Personality and Social Psychology, 55*(4), 580–587. doi:10.1037/0022-3514.55.4.580.

Fillmore, C. J. (1979). On fluency. In D. Kempler & W. S. Y. Wang (Eds.), *Individual differences in language ability and language behavior* (pp. 85–102). New York: Academic Press.

Forbes, R. J., & Jackson, P. R. (1980). Non-verbal behaviour and the outcome of selection interviews. *Journal of Occupational Psychology, 53*, 65–72.

Funes, K., & Odobez, J.-M. (2012). *Gaze estimation from multimodal kinect data.* Paper presented at the IEEE conference in computer vision and pattern recognition, Providence, RI, USA.

Gatica-Perez, D., Guillaume, L., Odobez, J.-M., & McCowan, I. (2007). Audio-visual tracking of multiple speakers in meetings. *IEEE Transaction on Audio Speech, and Language Processing, 15*, 601–616.

Guerrero, L. K. (2005). Observer ratings of nonverbal involvement and immediacy. In V. Manusov (Ed.), *The sourcebook of nonverbal measures.* Mahwah, NJ: Lawrence Erlbaum Associates.

Hollandsworth, J. G., Kazelskis, R., Stevens, J., & Dressel, M. E. (1979). Relative contributions of verbal, articulative, and nonverbal communication to employment decisions in the job interview setting. *Personnel Psychology, 32*, 359–367.

Imada, A. S., & Hakel, M. D. (1977). Influence of nonverbal communication and rater proximity on impressions and decisions in simulated employment interviews. *Journal of Applied Psychology, 62*, 295–300.

Janz, T. (1982). Initial comparisons of patterned behavior description interviews versus unstructured interviews. *Journal of Applied Psychology, 67*, 577–580. doi:10.1037/0021-9010.67.5.577.

Junker, H., Amft, O., Lukowicz, P., & Tröster, G. (2008). Gesture spotting with body-worn inertial sensors to detect user activities. *Pattern Recognition, 41*(6), 2010–2024. doi:10.1016/j.patcog.2007.11.016.

Kleinke, C. L. (1986). Gaze and eye contact: A research review. *Psychological Bulletin, 100*, 78–100.

Knapp, M. L., & Hall, J. A. (2010). *Nonverbal communication in human interaction* (7th ed.). Wadsworth: Cengage Learning.

Lathoud, G., & McCowan, I. A. (2003). *Location based speaker segmentation.* Paper presented at the meeting of acoustics, speech, and signal processing.

Leathers, D. G. (1992). *Successful nonvebal communication*. London: Collier Macmillan.

Leigh, T. W., & Summers, J. O. (2002). An initial evaluation of industrial buyers' impressions of salespersons' nonverbal cues. *Journal of Personal Selling & Sales Management, 22*, 41–53.

Lu, H., Rabbi, M., Chittaranjan, G. T., Frauendorfer, D., Schmid Mast, M., Campbell, A. T., & Choudhury, T. (2012). *Stresssense: Detecting stress in unconstrained acoustic environments using smartphones*. Paper presented at the UbiComp, Pittsburgh, USA.

Marcos-Ramiro, A., Pizarro-Perez, D., Marron-Romera, M., Nguyen, L. S., & Gatica-Perez, D. (2013). *Body communication cue extraction for conversational analysis*. Paper presented at the IEEE international conference on automatic face and gesture recognition, Shanghai, China.

McGovern, T. V., & Tinsley, H. E. A. (1978). Interviewer evaluations of interviewee nonverbal behavior. *Journal of Vocational Behavior, 13*, 163–171. doi:10.1016/0001-8791(78)90041-6.

Mehrabian, A. (1972). *Nonverbal communication*. New Brunswick: AldineTransaction.

Motowidlo, S. J., Carter, G. W., Dunnette, M. D., Tippins, N., Werner, S., Burnett, J. R., et al. (1992). Studies of the structured behavioral interview. *Journal of Applied Psychology, 77*, 571–587. doi:10.1037/0021-9010.77.5.571.

Nguyen, L. S., Odobez, J.-M., & Gatica-Perez, D. (2012). *Using self-context for multimodal detection of head nods in face-to-face interactions*. Paper presented at the international conference on multimodal interactions, Santa Monica, CA, USA.

Peterson, R. T. (2005). An examination of the relative effectiveness of training in nonverbal communication: Personal selling implications. *Journal of Marketing Education, 27*, 143–150. doi:10.1177/0273475305276627.

Pudil, P., Novovicova, J., & Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters, 15*, 1119–1125.

Quené, H. (2007). On the just noticeable difference for tempo in speech. *Journal of Phonetics, 35*, 353–362.

Reis, H. T., & Charles, M. J. (2000). *Handbook of research methods in social and personality psychology*. Cambridge: Cambridge University Press.

Taute, H. A., Heiser, R. S., & McArthur, D. N. (2011). The effect of nonverbal signals on student role-play evaluations. *Journal of Marketing Education, 33*, 28–40. doi:10.1177/0273475310389153.

Tiedens, L. Z., & Fragale, A. R. (2003). Power moves: Complementarity in dominant and submissive nonverbal behavior. *Journal of Personality and Social Psychology, 84*(3), 558–568.

Vinciarelli, A., Pantic, M., Bourlard, H., & Pentland, A. (2008). *Social signals, their function, and automatic analysis: A survey.* Paper presented at the ICMI 2008, Chania, Crete, Greece.

Wardhaugh, R. (1985). *How conversation works*. New York: Basil Blackwell.

Wood, J. A. (2006). NLP revisited: Nonverbal communications and signals of trustworthiness. *Journal of Personal Selling & Sales Management, 26*, 197–204. doi:10.2753/pss0885-3134260206.

Woolbert, C. (1920). The effects of various modes of public reading. *Journal of Applied Psychology, 16*, 162–185.